

Received April 13, 2022, accepted May 14, 2022, date of publication May 18, 2022, date of current version May 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176437

Large-Scale Assessment of Mobile Crowdsensed Data: A Case Study

CHRISTEL SIROCCHI¹, LORENZ CUNO KLOPFENSTEIN^{1,2},
AND ALESSANDRO BOGLIOLO^{1,2}

¹Dipartimento di Scienze Pure e Applicate, Università degli Studi di Urbino, 61029 Urbino, Italy

²DIGIT srl, 61029 Urbino, Italy

Corresponding author: Christel Sirocchi (c.sirocchi2@campus.uniurb.it)

ABSTRACT Mobile crowdsensing (MCS) is a well-established paradigm that leverages mobile devices' ubiquitous nature and processing capabilities for large-scale data collection to monitor phenomena of common interest. Crowd-powered data collection is significantly faster and more cost-effective than traditional methods. However, it poses challenges in assessing the accuracy and extracting information from large volumes of user-generated data. SmartRoadSense (SRS) is an MCS technology that utilises sensors embedded in mobile phones to monitor the quality of road surfaces by computing a crowdsensed road roughness index (referred to as P_{PE}). The present work performs statistical modelling of P_{PE} to analyse its distribution across the road network and elucidate how it can be efficiently analysed and interpreted. Joint statistical analysis of open datasets is then carried out to investigate the effect of both internal and external road features on P_{PE} . Several road properties affecting P_{PE} as predicted are identified, providing evidence that SRS can be effectively applied to assess road quality conditions. Finally, the effect of road category and the speed limit on the mean and standard deviation of P_{PE} is evaluated, incorporating previous results on the relationship between vehicle speed and P_{PE} . These results enable more effective and confident use of the SRS platform and its data to help inform road construction and renovation decisions, especially where a lack of resources limits the use of conventional approaches. The work also exemplifies how crowdsensing technologies can benefit from open data integration and highlights the importance of making coherent, comprehensive, and well-structured open datasets available to the public.

INDEX TERMS Collective intelligence, large-scale assessment, mobile crowdsensing, open data.

I. INTRODUCTION

Mobile crowdsensing (MCS) is a well-established paradigm of large-scale data collection for measuring and mapping phenomena of common interest for society. MCS applications are generally deployed on personal mobile devices and use the device's sensors and processing capabilities to measure, compute, and provide data to an application server, where collected data is further processed to gain insights and make predictions on the phenomenon under study. MCS has enabled a broad range of applications, including environment quality monitoring, noise pollution assessment, traffic planning, and public safety [1]. MCS applications have also been actively developed in road quality monitoring because of the road infrastructure's pivotal role in a country's socio-economic

The associate editor coordinating the review of this manuscript and approving it for publication was Giovanni Merlino¹.

development and the inability of traditional methods to provide a network-wide, frequent and reliable assessment of the road conditions.

Inadequate road infrastructure affects society in ways beyond economic considerations, especially in rural areas and developing countries, where it can limit the population's access to water, medical services, and education [2]. Poor road conditions are known to increase the fuel consumption and emissions of vehicles and adversely affect their suspension systems and other mechanical components [3]. Several studies have also demonstrated the influence of road-pavement surface conditions on driving safety and comfort [4].

Estimating a road lifespan is challenging because internal factors, namely materials used and construction methods, and external factors, such as weather conditions and traffic volume, can significantly affect the road integrity, sometimes

unpredictably. Timely and regular interventions can extend the life of a road by several years and drastically reduce its overall servicing cost because maintenance becomes more expensive as the quality of the road declines. Hence, being able to monitor road quality across an entire network and identify road sections that require urgent maintenance would result in lower costs, reduced pollution, increased safety and comfort for drivers [4].

Informed maintenance decisions need evaluating road conditions in a reproducible manner and on a sufficiently frequent basis across the entire network. Most road agencies do not dispose of sufficient resources and tend to prioritise maintenance based on expected economic returns [5]. Hence, preventive maintenance actions are disproportionately directed toward urban districts rather than rural areas. Accordingly, several low-cost solutions based on the MCS paradigm have been developed to address this lack of resources.

SmartRoadSense (SRS) is an MCS that utilises sensors embedded in mobile phones to monitor the quality of road surfaces by calculating a road roughness index (also referred to as P_{PE}). The SRS automatic and continuous road-monitoring system provides a regularly updated and detailed picture of the surface conditions. The open availability of SRS data is a significant resource for public managers of road infrastructures for road monitoring and maintenance planning. The main objection to the SRS methodology has been the lack of a clear relationship with more established indices, such as IRI and PCI. However, while weak correlations are likely to appear, a formula to convert among indices with precision is not to be expected. Other studies attempting to calculate such relationships, for instance, between IRI and PCI, rarely achieved conclusive results because of the radically different ways they are calculated. Thus, alternative methods are needed to test the validity of newly introduced road roughness metrics without relying on comparisons with other indices.

Crowd-powered data collection in MCS is significantly faster and more cost-effective than traditional methods. Still, it is inherently subject to more sources of variability, requiring strategies to assess and improve the accuracy of collected data. MSC's validation strategies are generally performed on a small scale and in a controlled environment, so they cannot evaluate the quality of massive datasets. Methods aimed at improving accuracy generally focus on outlier identification and filtering, inconsistency resolution, or reputation systems to identify improper contributions; hence they mainly address the accuracy problem at the level of the measurement, user or device. These methods also suffer scalability and computational cost limitations and cannot be systematically implemented.

This study addresses the need for efficient and versatile approaches to analysing and validating large crowdsourced data datasets. Firstly, statistical modelling of P_{PE} by the log-normal distribution under the law of proportional effects offers valuable insights on how to analyse and interpret this type of data correctly. Secondly, the integration of SRS data

with independently-sourced open datasets provides a statistical description of P_{PE} across the Italian road network and characterises the effect of various road qualities on the measured P_{PE} , which can help inform road construction and renovation decisions. Thirdly, the effectiveness of SRS is evaluated using known relationships between road features and pavement roughness. The adopted approach groups road sections in classes of predicted performance based on road features and confirms that the measured P_{PE} agrees with those assumptions, essentially replicating on a large scale what previous validation experiments have accomplished at a local level. It only addresses accuracy at the aggregate level, so its complexity, unlike most available methods, is independent of the number of user contributions. Finally, the work highlights the value of data integration in a crowdsensing platform and the importance of making data available to the public in a coherent, comprehensive and well-structured manner. Overall, the presented work aims at removing the obstacles to the correct and systematic use of SRS and other open crowdsensing platforms.

The remainder of the article is structured as follows: Section II revises relevant literature related to the topic, Section III outlines the proposed approach and describes the datasets analysed, Section IV presents and discusses the analysis results, and finally, Section V draws conclusions and discusses future work.

II. BACKGROUND AND PREVIOUS WORK

A. ROAD MONITORING TECHNOLOGIES

Pavement Roughness is widely considered the most suitable indicator of road condition and quality of travel. A traditional method based on visual inspections quantifies road roughness in terms of a pavement condition index (PCI). The U.S. Army Corps of Engineers developed this index as a numerical indicator of the surface condition of asphalt and concrete pavements with a value range of 0–100 [6]. PCI does not reflect structural properties of the road (e.g. capacity, resistance or roughness) but provides an objective basis for setting maintenance and repair requirements and priorities.

Previous studies have modelled the road elevation profile with impulsive functions, triangular waves, and sinusoidal signals [7], as the sum of randomly generated sinusoidal functions [8], or using stochastic models with low-pass characteristics [9]. Based on this work, systems gauging the road roughness by taking physical measurements of the road irregularities have been actively developed in the following decades. These methods fall into three main categories: i) methods using laser measurements, ii) methods calculating vehicle vibration, iii) methods applying image recognition techniques [10].

Laser measurement-based detection is a consolidated approach that entails adopting sophisticated equipment installed on separate inspection vehicles, such as laser profilers and data acquisition systems, to convert the road surface into a three-dimensional object in a coordinate

system [11], [12]. These methods use the “international roughness index” (IRI), proposed by Sayers *et al.* [13] and adopted worldwide as the standard index for road roughness estimation. Calculating IRI is challenging because it considers the vehicle’s characteristics in addition to the conditions of the road. Laser measurement-based approaches are still considered the most accurate in evaluating the road-surface condition. However, they pose a considerable expense (including machinery, installation and calibration) and generally cannot process road data in real-time.

Vibration-based methods detect road-surface anomalies by measuring the vehicle oscillations using autonomous accelerometers or accelerometers incorporated in mobile devices. The rapid development of mobile sensing technology has empowered vehicles owned by the general public with the ability to collect vibration measurements and contribute to road monitoring. Erikson *et al.* [14] introduced a system named the “Pothole Patrol” that used a set of high-frequency accelerometers and Global Positioning System (GPS) receivers deployed in embedded computers in vehicles. Mohan *et al.* [15] introduced “Nericell”, the first road and traffic monitoring system using sensors found in smartphones. “SmartRoadSense” [16] embraces a similar technical approach, utilising accelerometers and GPS receivers in mobile devices, but has the broader scope of quantifying road quality across the entire road network. Chen *et al.* [17] used hardware modules equipped with low-end accelerometers and GPS receivers mounted on distributed vehicles, together with a lightweight data mining algorithm to detect road potholes. Yi *et al.* [18] proposed a signal processing technique to calculate the vertical component (VC) of acceleration and a sensing algorithm to detect potholes and bumps based on VC. Mohamed *et al.* [19] proposed a road condition monitoring framework that senses road anomalies using the gyroscope around gravity rotation combined with the accelerometer as a cross-validation method. Bhatt *et al.* [20] developed a mobile application that uses a support vector machine (SVM) classifier to detect potholes and assess road conditions in real-time. Allouch *et al.* [21] developed the “RoadSense”, which applies a decision tree classifier to tri-axial accelerometer and gyroscope data from mobile devices to predict road quality automatically. The method proposed by Jang *et al.* [22] comprises a vehicle client which measures vibrations and performs a preliminary defect detection and a back-end server collecting and analysing data by supervised machine learning technique and a trajectory clustering algorithm. Nunes and Mota [23] proposed a framework for participatory road monitoring named “Streetcheck” that gathers sensor data from mobile devices together with users’ ratings on road surface quality and employs supervised learning algorithms to classify road surface quality. All vibration-based approaches require filtering steps to extract only vibrations associated with road anomalies. They are suitable for real-time evaluation of road-surface conditions at a reduced cost. However, they cannot assess pavement anomalies in

areas other than the vehicle wheel paths, thus requiring a large number of measurements to quantify the road-surface damage. Moreover, due to data privacy laws, most crowdsourced vehicular vibration data do not contain user information, such as vehicle type, model, physical parameters; hence traditional IRI-estimation models cannot be directly applied.

Finally, Image Recognition (IR)-based detection techniques use IR methods, such as deep neural networks (DNNs), to capture images and detect road damages in real-time. This approach has recently attracted attention as it can analyse road-surface conditions over a wide area at a reasonable cost. However, large and accurately labelled image datasets of the road-surface conditions are required to train a DNN, and the effect of various factors (e.g. road-surface colour, illumination, weather,..) on recognition rate is still unclear [10].

B. SmartRoadSense

This section provides an overview of the work conducted within the SRS project. It is structured as follows: the signal processing algorithm employed and the mathematical model upon which it was developed are described in II-B1, the system architecture and data aggregation techniques are outlined in II-B2, validation experiments are summarised in II-B3, and previous studies of the relationship between the road roughness index and the vehicle speed are shown in II-B4.

1) MATHEMATICAL MODEL

Gillespie [8] showed that the typical Power Spectral Density (PSD) of the road vertical profile had a low-pass characteristic that decreased with increasing spatial frequency. Thus, the road surface could be modelled as a white Gaussian noise filtered by a first-order low-pass filter. The white gaussian noise has autocorrelation function $p_{ww} = q\delta(x)$, where q is the PSD magnitude, and $\delta(x)$ is the Dirac delta function. PSD is defined as $S_{ww}(\Lambda) = q$, where Λ is the spatial frequency expressed in cycles. The frequency response of the low-pass filter is

$$H(\Lambda) = \frac{1}{p + j2\pi\Lambda} \quad (1)$$

so that the PSD of the road elevation profile $S_{rr}(\Lambda)$ becomes

$$S_{rr}(\Lambda) = S_{ww}(\Lambda)|H(\Lambda)|^2 = q\left|\frac{1}{p + j2\pi\Lambda}\right|^2 \quad (2)$$

where the statistical properties of the road profile are completely characterised by parameters p and q . From equation 2 it can be derived that the same p and q road parameters can be obtained by analysing the PSD of the vertical acceleration of a point closely following the road profile [8], [24]. In real applications, an accelerometer senses the road through tires, suspensions, and the mechanical coupling to the vehicle, combined with the accelerations provided by gravity, vehicle speed changes, centrifugal acceleration at curves, and engine vibrations. Thus, the accelerometer senses and samples the PSD of the acceleration filtered by an unknown transfer

function modelling the effect of undesired contributions, producing a discrete-time vector signal composed of the triaxial components. Because most of these unwanted accelerations have a constant or periodic spectral content, a prediction filter can remove them [25], isolating accelerations attributable to irregularities in the road surface. SRS utilises Linear Predictive Coding (LPC) analysis, a signal processing technique that estimates the current value of a sample $a(n)$ as a linear function of its past values [26]:

$$e(n) = a(n) \sum_{i=1}^N \lambda_i a(n-i), \quad (3)$$

where N is the filter memory length, λ_i with $i = 1, \dots, N$ are the LPC coefficients, and $e(n)$ is the residual prediction error. The prediction filter is computed with the Levinson-Durbin recursion, and the error is calculated on signal segments of length M . The prediction error $e(n)$ retains the information on the road parameter q , while the parameter p is filtered out. A parameter proportional to q is obtained by estimating the power of the prediction error (P_{PE}) on each segment:

$$P_{PE} = \frac{1}{N} \sum_{n=1}^{M-1} e(n)^2. \quad (4)$$

Finally, an index of the road roughness is calculated by averaging the P_{PE} for the three axial components:

$$P_{PE} = \frac{1}{3} (P_{PE_x} + P_{PE_y} + P_{PE_z}). \quad (5)$$

2) SYSTEM ARCHITECTURE

The SmartRoadSense architecture comprises three major components:

- a mobile application at the user level that processes raw data from the embedded accelerometers and transmits the result and the geographic localisation data to a server;
- a cloud-based back-end where georeferenced data are mapped, aggregated, and stored;
- a web-based graphical front-end for data visualisation [16].

SmartRoadSense employs mechanisms of reduction and aggregation of geo-localised data to mitigate the impact of the large volumes of data continuously produced by the sensing devices, facilitating data analysis and visualisation while retaining relevant information. Spatial aggregation is performed in the cloud by mapping the points received by the back-end onto a map database and aggregating them according to given spatial constraints. Data falling within an area of fixed radius are reduced to a single value, smoothing the effect of possible outliers. Temporal aggregation of these quantities is achieved by periodically calculating the weighted average of points over time, incrementally down-weighting older data [27]. SmartRoadSense also implements efficient algorithms that detect and correct mapping artefacts and significantly enhance the accuracy of

map-matching of user-supplied geospatial data in crowdsensing applications [28].

3) VALIDATION APPROACHES

Several approaches for data validation have been adopted during the development of SRS to verify the quality of the collected data and, when possible, its correlation with ground-truth data [2]. First, visual inspection data gathered by road maintenance experts were compared with the SRS aggregated data on regional roads. Then, defect reports and visual inspection data on selected sections of the highway network were compared to the corresponding SRS. Finally, in the Mantova pilot, a road technician visually inspected the complete network of the municipal area, annotating significant events, while a software application recorded pictures of the road, and SRS calculated the road roughness index [2].

A recent study compared SRS data mapped on Provincial Road 2 (SP2) in Salerno with the Distress Cadastre data and PCI assessments for the same highway and found that, although the effectiveness varies with the distress type, SRS is efficient in monitoring the most critical road failures [29]. In particular, SRS proved accurate in identifying distresses characterised by vertical thicknesses while less sensitive toward superficial damages. However, an exact correspondence between the two indices is not to be expected given the different ways in which they are calculated: PCI is based on inspection and better characterises visible damages, while P_{PE} is prone to filter out constant roughness but is better equipped to detect unforeseen road events which might not be visible.

4) VEHICLE SPEED

A study aimed at modelling the influence of vehicle speed on the measurement suggested that the SRS roughness index depends to a certain degree on the vehicle speed. Controlled studies showed that the value of P_{PE} computed on a single device attached to a specific vehicle travelling on the same road depends on the speed of the car according to a gamma law, with parameter $\gamma \in [0,2]$ and with γ tending to decrease for increasing values of the vehicle speed. Performing the same analysis on data grouped by type of road (motorway, trunk primary, secondary), they found that each cluster showed a degree of dependency of the P_{PE} from the vehicle speed that could be modelled with a gamma law [30].

C. ASSESSING QUALITY OF CROWDSENSED DATA

MCS systems require the participation of a multitude of users to be effective. Scalability is always a key factor to consider when designing solutions for MCS platforms, as they are expected to handle large amounts of data [31]. The ubiquitous and open nature of the crowdsensing paradigm allows for the fast and cost-effective collection of a vast amount of data but also exposes the system to malicious or erroneous contributions. Hence, mechanisms to evaluate data trustworthiness are required to guarantee the quality and reliability of service. Three main types of deleterious data affect MCSs,

which require individual consideration and the appropriate corrective action: missing data (e.g. data lost or not volunteered), unreliable data (e.g. due to noise or faulty sensors), and manipulated data (e.g. to increase the user's data utility in the system). The problem of ensuring trustworthiness in MCSs is of great importance but remains open [31].

Missing data can be detected and corrected using compressive sensing, a technology that reconstructs a signal and recovers incomplete data sets by sampling sparse signals under the sub-Nyquist rate and applying computationally intensive algorithms [32]. This technique requires that multiple observations exist and that trusted users would never contribute faulty data [33]. Quality of contributed data can be evaluated by modelling evidence from (i) ground truth, (ii) similarity-based outlier detection, (iii) prior reputation context, and (v) rating feedback mechanism [34]. Ground truth data is often unavailable, and acquiring it entails investing dedicated resources, invalidating the benefits of crowdsensing. Similarity-based methods filter improper contribution by awarding higher trust to measures that are closest to data already collected [35]. However, multiple similar erroneous contributions, either intentional (e.g. Sybil attacks) or non-intentional (e.g. sensor defect), can impair future quality assessments. Reputation methods assign a trust score based on previous user activity but assume the existence of prior and reliable reputation scores [35]. They are particularly suitable to be applied in social participatory sensing platforms, but the complexity of the reputation algorithm, such as PageRank, is at least as high as the number of nodes and edges of the network, limiting the size of the social network and the scalability of the approach. Additionally, changes in user behaviour and cold start, i.e. lack of initial reputation values, are not adequately addressed. In rating feedback mechanisms, other agents evaluate the information contributed by the users. This approach is simple, fast and cost-effective but is susceptible to other threats and is not suitable to assess certain types of data, such as sensor data [34].

A versatile method for estimating crowdsensed data accuracy which does not make assumptions on the statistical distribution nor requires additional information (e.g. ground truth data or reputation scores) combines statistical bootstrap with uncertainty propagation [36]. This method was validated on a short section of SmartRoadSense data but cannot be systematically implemented due to the algorithm's complexity, which resamples at least 10^4 times in each sample at each cycle. In conclusion, the discussed methods raise concerns regarding scalability and applicability, motivating the search for alternative approaches.

III. VALIDATION APPROACH

A. OPEN DATA INTEGRATION

This study investigates the effect of various aspects of a road setting, such as construction materials, population density, presence of a bridge, road ranking and speed limit, on the value of P_{PE} measured on that road. The effect on road quality and performance is known for some of these features and can

be used to test SRS's ability to capture road performance, hence evaluating the overall trustworthiness and truthfulness of the data under study. The adopted approach integrates data from multiple open datasets to enrich the original dataset with additional features, then isolates roads with characteristics commonly associated with higher surface roughness and verifies that P_{PE} computed on those roads agrees with these assumptions.

The methodology applied is, in fact, a large scale version of that employed to validate the SRS method, where a series of experiments classified road sections into performance groups and compared road roughness metrics calculated for each group [2]. However, these experiments were expensive and time-consuming, as they relied on specialised software, visual inspections, geo-referenced annotations and the feedback of road technicians and professionals, thus not applicable on a large scale. Conversely, the proposed approach is scalable and versatile, as it relies on readily available independent open data. It can be readily applied to any crowdsensed dataset for which reasonable assumptions can be made and constitutes a valid alternative when collecting additional data, employing specialised equipment or consulting experts is not applicable. However, the method's applicability and efficacy strongly depend on the availability of data related to the phenomenon under study, highlighting the importance of open data.

The chosen approach only addresses accuracy at the aggregate level, whereas most available methods focus on individual measurements, devices, locations or users. It is argued here that, in crowdsourcing systems, accuracy should only be relevant at the aggregated level. Crowdsourced data is inherently prone to variability and cannot reach the accuracy standards of controlled experiments. On the other hand, this is not necessary because the "wisdom of the crowd" principle [37] exploited in this paradigm attests that the cumulative effect of all measures can capture the underlying information even where the individual measurement fails to do so [38].

B. OPEN DATA SETS

The SRS dataset, whose structure is outlined in Table 1, was retrieved from the SmartRoadSense project Open Data section in March 2021 [39]. The vast majority of points are mapped in Italy (75.1% points), motivating the choice to limit the analysis to the Italian road network. Values of P_{PE} below 0.001 or above 4.0 were removed from the dataset as recommended in [16].

Boundaries of Italian territories, population and urbanisation level data for each Italian city were obtained from the Italian National Institute of Statistics (ISTAT) [40] and added to the SRS dataset. The urbanisation level measures demographic density calculated using one sq. km grids and can assume three values (1,2,3) expressing decreasing population density.

OpenStreetMap (OSM) data mapped on the Italian territory was downloaded in PFB format from the Geofabrik server [41]. Relevant OSM tags were copied using an OSMHandler based on Osmium (a data processing Python

TABLE 1. SmartRoadSense dataset structure.

Attribute	Description
latitude, longitude	spatial coordinate in Coordinate Reference System (CRS) "EPSG:4326"
ppe	aggregate road roughness index calculated by the SRS pipeline (P_{PE})
osm_id	ID number identifying the road the point is mapped onto in the OpenStreetMap database
highway	OSM "highway" tag of the road (see Table 4)
quality	index of data accuracy
passengers	number of vehicle passengers
updated_at	timestamp of the latest data aggregation for that point (the aggregated value is updated only if new data was mapped in the area afferent to that point in the last epoch)

library for OSM data) and added to the SRS data set. Given the many tag values associated with only a handful of points in the SRS dataset, only "surface" and "bridge" tag values with at least 100 points were considered. Similarly, only the most frequent speed limits were included in the analysis: 90, 110 and 130 for motorways, 30, 50, 70, 90 for the other roads. Unspecific "surface" tags such as "paved," "unpaved", and "gravel" were also excluded.

IV. RESULTS

The results of the investigation are presented as follows: Section IV-A offers theoretical and empirical evidence to the claim that P_{PE} is best modelled by the log-normal distribution, Section IV-B describes and discusses the relationship between a variety of road feature and the corresponding P_{PE} , and finally, Section IV-C evaluates the coherence of these results with previous work on the relationship between P_{PE} and vehicle speed.

A. STATISTICAL MODELLING OF P_{PE}

1) THEORETICAL CONSIDERATION

An adequate description of the variability of a quantity is a prerequisite for studying its patterns and estimating variance components. The Gaussian distribution is generally the first choice when modelling a real-valued variable affected by random variation, but it is not the most suitable distribution in a variety of circumstances [42]. Arithmetic mean and standard deviation are unsuitable statistics for skewed distributions, which are better characterised by the geometric counterpart or in terms of a transformed variable that distributes normally. For right-skewed variables that only take positive values, the log-normal distribution fits the data better than the normal, with the advantage of linking back to a normal distribution applying a simple logarithmic transformation.

A random variable X is said to be log-normally distributed if $\log(X)$ is normally distributed. Let Z be a standard normal variable and μ , and σ be two real numbers. Then, the distribution of the random variable $X = \exp(\mu + \sigma * Z)$ is the log-normal distribution with parameter μ and σ . The probability density function for the log-normal

distribution is:

$$p(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{6}$$

where μ is the mean and σ is the standard deviation of the normally distributed logarithm of the variable. A log-normal distribution is often specified in terms of its geometric mean μ^* and standard deviation σ^* , which can be easily computed from μ and σ as

$$\mu^* = e^\mu, \quad \sigma^* = e^\sigma \tag{7}$$

Notably, μ^* is also the median of the distribution since μ is the median of the transformed variable. Fig. 1 exemplifies how the parameter σ^* defines the shape of the distribution, while the median μ^* affects the horizontal and vertical scaling leaving the shape unchanged.

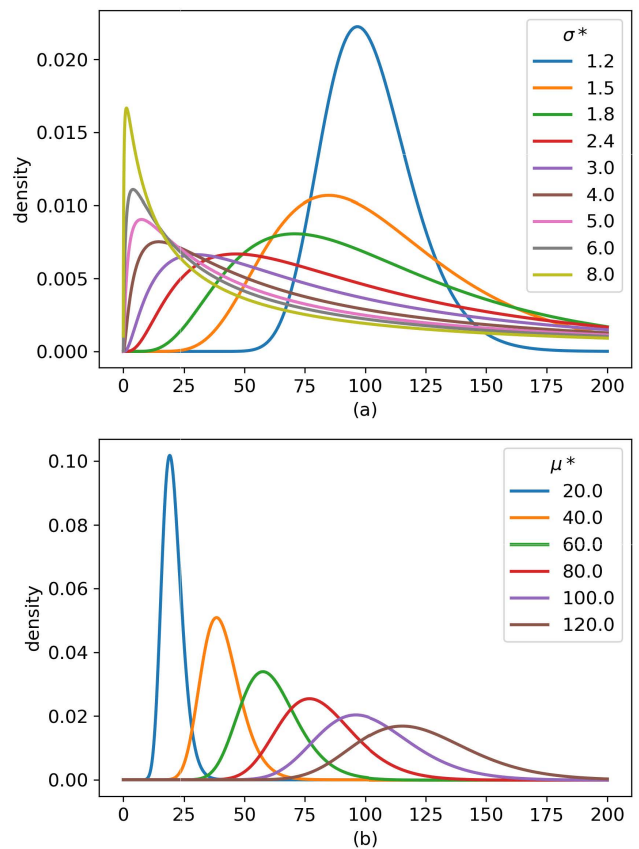


FIGURE 1. Density functions of log-normal distributions having (a) $\mu^* = 100$ and different values σ^* , (b) $\sigma = 0.2$ and different values of μ^* .

Log-normally distributed data are frequently, but incorrectly, described in terms of the arithmetic mean \bar{x} and standard deviation s . Estimates of their geometric counterparts can be obtained from equations (8) so that more informative statistics can be derived once the log-normal trend is identified [43]. Moreover, the geometric mean is always less than or equal to the arithmetic mean due to the AM–GM inequality and the logarithm being a concave function. Failing to recognise the log-normal trend could overestimate the sample's

central value.

$$\mu^* = \frac{\bar{x}}{1 + \frac{s^2}{\bar{x}}}, \quad \sigma^* = \exp \sqrt{\ln 1 + \frac{s^2}{\bar{x}}} \quad (8)$$

The Central Limit Theorem in the log domain states that a log-normal process is the statistical realisation of the multiplicative product of many independent positive random variables, as opposed to its primary formulation, where the sum of many independent identically distributed random variables approximates the normal distribution. Thus, normal and log-normal distributions both describe forms of variability based on many forces acting independently of one another but of additive and multiplicative nature, respectively.

The log-normal distribution may also be seen as a particular outcome of Gibrat’s law, also known as the law of proportionate effect. In a growth process, Gibrat’s law states that the probability of a given growth rate for a particular entity is independent of its size, so that growth in proportion to size is a random variable [44]. Given an entity of size X_0 , at each step j , the entity may change in size according to a random variable F_j , so that

$$X_j = F_j X_{j-1}. \quad (9)$$

Considering the expression in the log domain,

$$\ln X_j = \ln X_0 + \sum_{k=1}^j \ln F_k \quad (10)$$

if the $\ln F_j$ are identically distributed random variable, according to the Central Limit Theorem $\sum_{k=1}^j \ln F_k$ converges to a normal distribution and for sufficiently large j , X_j approaches a log-normal distribution [45].

The log-normal model can be theoretically derived under assumptions matching several failure degradation processes, such as corrosion, diffusion, migration, crack growth, electromigration. It is frequently used to model failure of a fatigue-stress nature under the “multiplicative degradation argument”, which reformulates the law of proportional effect for failure and degradation processes [46].

2) EMPIRICAL EVIDENCE

The data under study already fulfil the main requirements for rejecting the normality assumption and considering the log-normal trend: it can only assume positive values and is right-skewed. The best-fit normal distribution for P_{PE} , calculated with Maximum Likelihood Estimator (MLE), assigns 99.7% of the values (3 standard deviations from the mean) in the range $[-0.629, 1.067]$ (Fig. 2 a, b), with a 20% probability assigned to negative values. The data is positively skewed, with skewness = 4.248. The arithmetic mean ($\mu = 0.219$) and standard deviation ($\sigma = 0.282$) of P_{PE} showed that $\mu \pm \sigma$ interval $[-0.063, 0.502]$ contains 90.6% of data instead of 68.2%, that 69.1% points fall below μ instead of 50%, and that data below the mean are found exclusively within one standard deviation (Fig. 2 c). Conversely, the $[\mu^*/\sigma^*, \mu^* \times \sigma^*]$ interval (1 standard deviation from

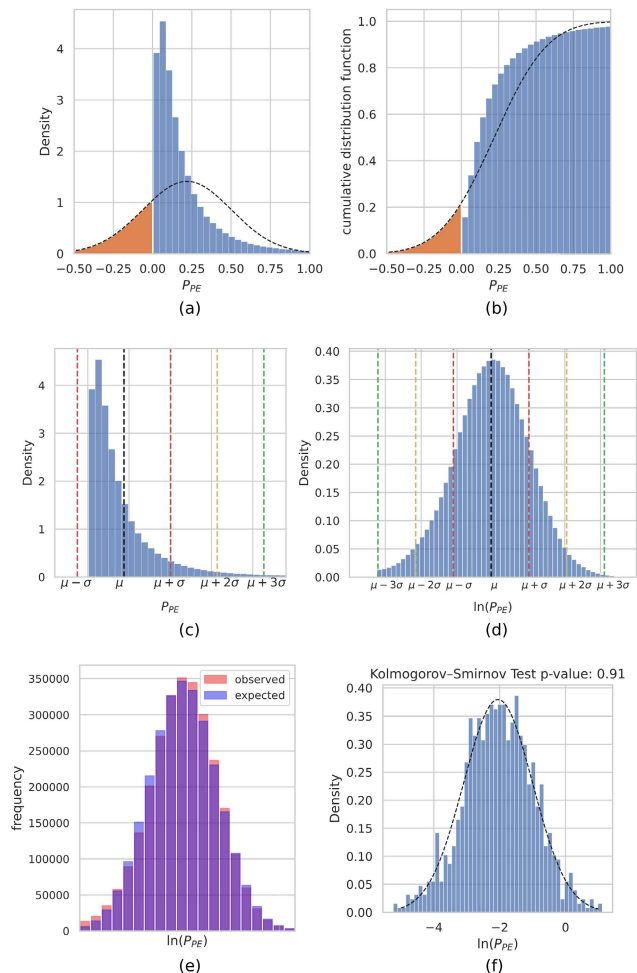


FIGURE 2. (a) Probability density function and (b) cumulative distribution function of P_{PE} and the corresponding best-fit normal distribution; (c) the arithmetic mean and standard deviation of P_{PE} , and (d) the geometric mean and standard deviation in the transformed domain; (e) $\ln(P_{PE})$ observed bin size compared to expected bin size, and (f) Kolmogorov-Smirnov test for normality.

the mean) contains 68.8% data, the $[\mu^* / (\sigma^*)^2, \mu^* \times (\sigma^*)^2]$ interval (2 standard deviations from the mean) contains 95.0%, the interval $[\mu^*/(\sigma^*)^3, \mu^* \times (\sigma^*)^3]$ contains 99.9%, and 49,2% points fall below μ^* , suggesting that the geometric mean and standard deviation are far more useful in characterising the data under study (Fig. 2 d). The log-transformed data is symmetrical and has skewness = -0.113 . The Kolmogorov–Smirnov test was used to compare random samples from the transformed data with the best fit normal distribution. It consistently returned p-values greater than 0.05, i.e. the null hypothesis that the two distributions are identical cannot be rejected (Fig. 2 e, f). These results were also confirmed in road-specific data.

The log-normal nature of SRS data can be seen as the realisation of the law of proportionate effect. Under this assumption, the road roughness index measured by SRS results from the multiplicative effect of many independent positive random variables. This index, as an indicator of the structural stability of the road pavement, can also be modelled

under the “multiplicative degradation argument”, meaning that, at any given time, the road degradation rate is independent of its current level of degradation but proportionate to it by a random factor. Notably, failure mechanisms like crack growth and propagation, which are primarily involved in the road degradation process, have already been modelled according to the law of proportionate effect [47]. This model is based on a stochastic process and is not concerned with specifying the effects of identifiable growth influences. In the study of city growth, Gibrat argued that the very nature of the problems under study is so complex that such an approach is appropriate. Similarly, the factors affecting the road surface conditions and their interactions are so complex that the model can be applied without the need to identify its underlying factors [44].

B. LARGE-SCALE ASSESSMENT OF SmartRoadSense DATA

This section investigates the relationship between road features (construction materials, population density, expansion joints, road rank and the speed limit) and the road roughness index computed by SRS. Assumptions on the effect of some of these road features were formulated based on results of small-scale validation experiments performed within the SRS project, the literature on road performance and the common understanding of the subject. The investigation showed complete coherence between the road roughness crowdsensed data collected through the SRS platform and the road structural data contributed in OSM by a multitude of users. Features that are known to increase road roughness, such as expansion joints and certain construction materials, were consistently associated with higher values of P_{PE} , supporting the claim that SRS can accurately detect road roughness on a large scale. The study of these and other features (e.g. road ranking and speed limit) offers insights into the pavement conditions across the entire network and can help inform road construction and renovation decisions.

1) PRESENCE OF A BRIDGE

Expansion joints are commonly found in bridges and viaducts to guarantee their structural integrity all year. These joints are expected to cause mechanical solicitations on the travelling vehicle, which are detected by the SRS system and reflected in a higher P_{PE} value. Roads passing on a bridge (such as viaducts and overpasses) were identified using the ‘bridge=*’ tag in OSM. Its most frequent tag values are listed in Table 2. As expected, the geometric mean of P_{PE} mapped on roads having the bridge tag was 20 to 35% higher than that of points mapped on other roads (Fig. 3 a). The independent T-test found these differences significant (p-value = 0.000) for each road type. Fig. 3 b compares in the transformed domain two same-sized samples of motorway points, the first from roads not tagged with the bridge feature and the second from bridge roads. The distribution of bridge points deviates from normality and has a skewness value three times higher than the other group. The Kolmogorov–Smirnov

TABLE 2. Most frequent tag values for “bridge” in OSM.

Key	Value	Description
bridge	yes	non-specific bridge tag, possibly combined with other bridge:* tags
bridge	aqueduct	a longer structure for carrying a canal
bridge	viaduct	a bridge composed of a series of spans (may be arches, girders supported by piers, etc.)

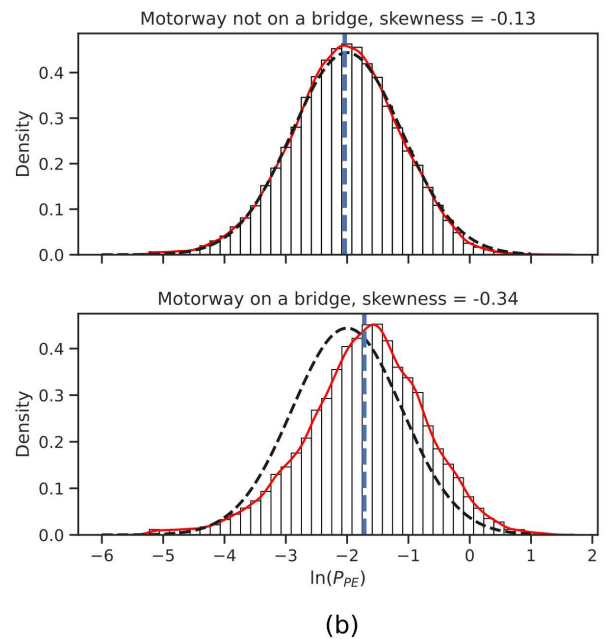
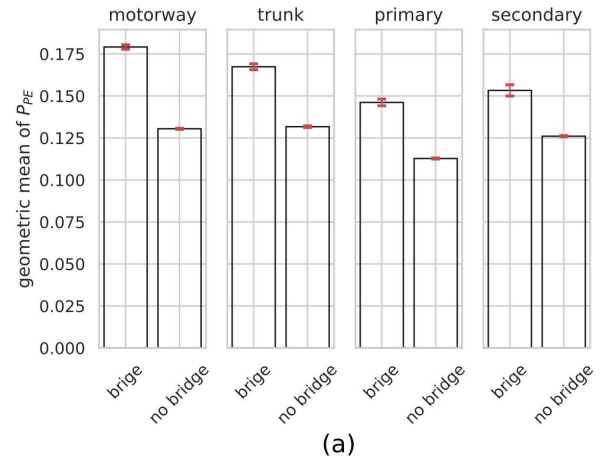


FIGURE 3. Comparison of (a) the geometric mean of P_{PE} and (b) the distribution of $\ln(P_{PE})$ in roads on a bridge and not on a bridge.

test applied to both samples found that the subset of bridge points does not belong to the best-fit normal distribution (p-value = 0.000). The loss of normality is coherent with the modelling of P_{PE} as the realisation of the Central Limit Theorem, according to which the contribution of each factor should be equal and infinitesimal to observe normality in the transformed domain.

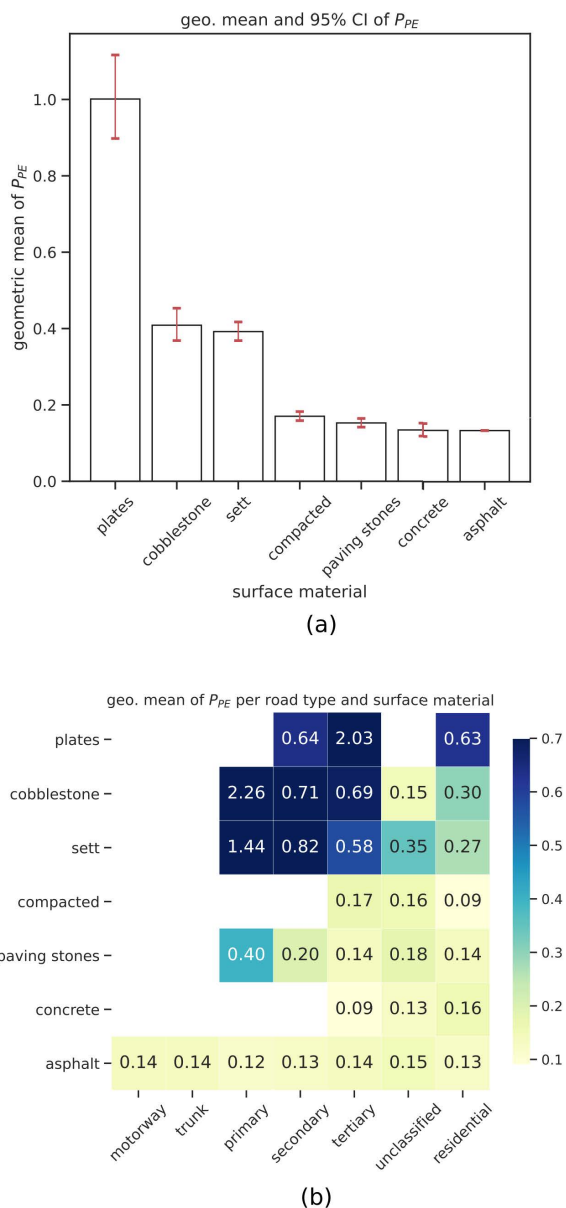


FIGURE 4. Geometric mean of P_{PE} with in roads grouped by (a) construction material (b) and also road category.

2) CONSTRUCTION MATERIALS

Materials and construction methods affect the smoothness, homogeneity, and overall quality of roads. Asphalt and concrete are the first two choices when building high-performance roads; hence they are expected to measure a lower P_{PE} compared to other materials. Compacted roads are considered the next best material for road construction because of their stability and grip and are expected to record a relatively good average P_{PE} . Similarly, roads made of large blocks separated by very narrow gaps (paving stones) should provide a lower P_{PE} compared to those made of smaller bricks with wider gaps (sett, cobblestone). SRS data were grouped based on the construction material of the

TABLE 3. Most frequent tag values for "surface" in OSM.

Key	Value	Description
surface	paved	a feature that is predominantly paved; i.e., it is covered with paving stones, concrete or bitumen (rough description)
surface	unpaved	a feature that is predominantly unsealed (unpaved); i.e., it has a loose covering ranging from compacted stone chippings to earth (rough description)
surface	gravel	very large meaning range, from huge gravel pieces like track ballast used as surface, through small pieces of gravel to compacted surface
surface	asphalt	short for asphalt concrete - mineral aggregate bound by asphalt (most such features are tagged as paved without specifying the exact surface)
surface	paving_stones	a relatively smooth surface paved with artificial blocks (block pavers, bricks) or natural stones (flagstones), with a flat top, with very narrow gaps between individual paving stones
surface	concrete	cement based concrete, forming a large surface, typically cast in place and may have predetermined breaking joints
surface	gravel	a mixture of larger (e.g., gravel) and smaller (e.g., sand) parts, compacted; best sort of ways below paving with asphalt, concrete, paving stones
surface	sett	sett paving, formed from natural stones cut usually to a regular shape (the stones do not cover the surface completely, unlike paving stones)
surface	cobblestones	generic value for cobblestone in the colloquial sense.
surface	concrete_plates	heavy-duty plates chained closely together on the short side (might have tar or sand in between the connections)

corresponding road using OSM "surface=*" tag, whose most relevant values are listed in Table 3. The geometric mean and 95% confidence interval were calculated for each group, confirming that P_{PE} positively correlates with the paving material performance. The most performing road paving materials, asphalt and concrete, had the lowest average P_{PE} , followed by paving stones, compacted (the best choice material after asphalt, concrete and paving stones according to OSM), sett and cobblestone, and finally concrete plates (Fig. 4). The non-parametric Kruskal-Wallis H-test found the means to be different (p-value = 0.000). Post hoc comparisons between groups showed that each pair of subsets is significantly different apart from sett + cobblestone (p-value = 0.498), concrete + paving stones (p-value = 0.085), asphalt + concrete (p-value = 0.966), confirming that materials with similar properties record comparable road roughness values. P_{PE} values tend to be consistent for a particular material across different road types. Asphalt roads, for example, produce consistently low P_{PE} values across the entire road network, while concrete plates are consistently associated with very high P_{PE} values (Fig. 4 b). These results suggest that SRS can detect material-related irregularities and quantify their magnitude.

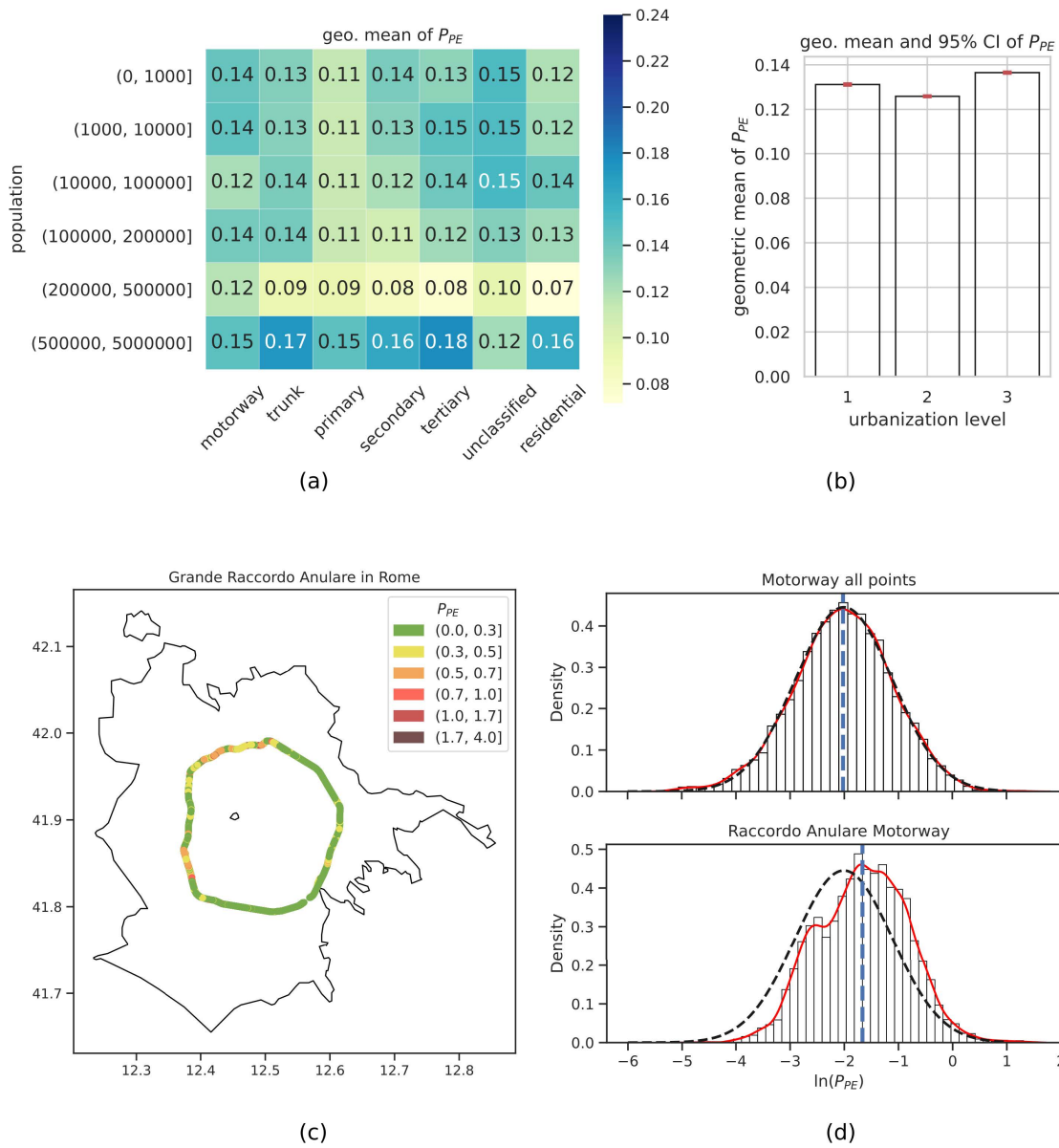


FIGURE 5. Geometric mean of P_{PE} per (a) population size and (b) urbanisation level of the afferent city. (c) P_{PE} measured in the most trafficked Italian highway and (d) its distribution in the transformed domain.

3) POPULATION DENSITY AND TRAFFIC FLOW

Roads affected by heavy traffic and frequent heavy trucks are expected to get damaged faster and are generally scheduled to receive maintenance every few years. Given the lack of annual average daily traffic data (AADT) for Italy or other network-wide open traffic datasets, population data from ISTAT was used to approximate the traffic flow. SRS data were binned according to the population and urbanisation level of the afferent city, and relevant statistics were calculated for each group. Cities with more than 0.5 million inhabitants, which generally have the role of regional capitals, have the highest P_{PE} values across all road types, indicating that densely populated areas have the lowest road quality. Conversely, medium-large cities, which are generally

province capitals, have the lowest P_{PE} values and the best road performance (Fig. 5 a), possibly because their essential administrative role grants them more frequent road improvements while suffering only moderate traffic flow. Similarly, cities having a medium urbanisation level recorded in average the lowest P_{PE} (Fig. 5 b). Independent T-test found the differences between groups to be significant.

The relationship between traffic flow and P_{PE} can be appreciated on the busiest motorway in Italy, the “Grande Raccordo Anulare” in Rome (OSM Relation: Grande Raccordo Anulare, osm_id: 1331370), which counts up to 165.000 vehicles per day according to ANAS, the National Autonomous Roads Corporation. The geometric mean calculated on this road is 0.17, 30% higher than that calculated on

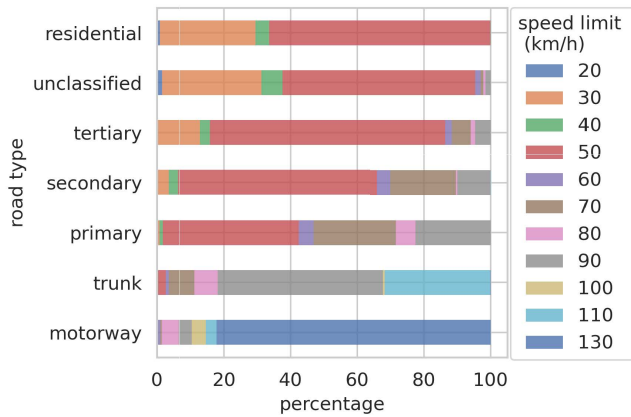


FIGURE 6. Distribution of speed limit per road ranking in the Italian road network.

all motorways (Fig. 5 c). The distribution of the transformed P_{PE} visually displays a degree of deviation from normality and differs significantly from the best-fit normal distribution for motorways (Fig. 5 d), suggesting predominant factors at play.

4) ROAD RANKING

The role of a road within a road network is its most defining feature. Motorways and trunks make up the backbone of the transport infrastructure and conduct most long-distance traffic. Primary, secondary, tertiary and unclassified roads lead medium-distance traffic and connect progressively smaller settlements, while residential roads provide access to housing within settlements. Each of these road categories is characterised by a speed limit, an average vehicle speed, a road type (urban, extra-urban or both), preferred construction materials, traffic intensity, standards of quality, etc. OSM defines a clear road hierarchy, outlined in Table 4, which is why this composite attribute is renamed here as “road ranking”. It is not possible to safely make assumptions on the effect of this attribute on P_{PE} due to several concurring factors. More important roads are built with better performing materials and according to higher quality requirements, so they are expected to record lower P_{PE} . However, vehicles travel on these roads at a much higher speed, a parameter known to affect road roughness measurement in all vibration-based methods. In Fig. 6, the speed limit distribution per road category shows how more important roads tend to have higher speed limits.

The road ranking does not show a direct correlation with the mean of the corresponding best-fit normal distribution (Fig. 7 a). However, when including the type parameter, the model could explain 89% of the variability in the data (R-squared = 0.89) with both variables having significant explanatory power (p-value < 0.05). In urban roads (primary, secondary, tertiary, unclassified, residential) with a speed limit of 50 Km/h and 40 km/h, a linear relationship with the mean is found in both groups with R-squared = 0.863, p-value = 0.022 at 50km/h speed limit (Fig. 7 b) and

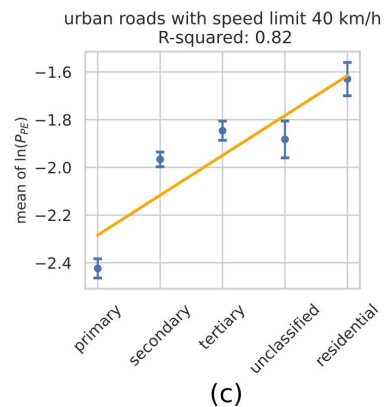
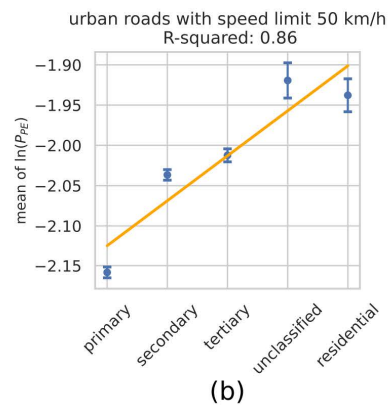
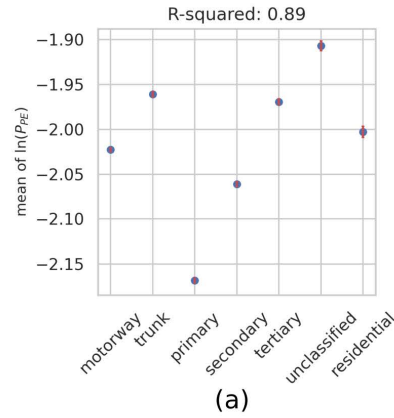


FIGURE 7. Relationship between road ranking and the mean of $\ln(P_{PE})$ in (a) all roads, (b) urban roads with speed limit of 50 km/h and (c) urban roads with speed limit of 40 km/h.

R-squared = 0.816, p-value = 0.036 at 40km/h speed limit (Fig. 7 c). Hence, P_{PE} correlates with the road ranking, as one would expect, when controlling for vehicle speed or road type. A strong linear correlation was found between the road ranking and the standard deviation of $\ln(P_{PE})$, with an R-squared value of 0.937 and p-value = 0.000 which persists within individual regions across the country (Fig. 8). This analysis was limited to urban roads, for which 50 km/h is the speed limit across all road categories 4, and 40 km/h is generally associated with good quality roads. It could not be

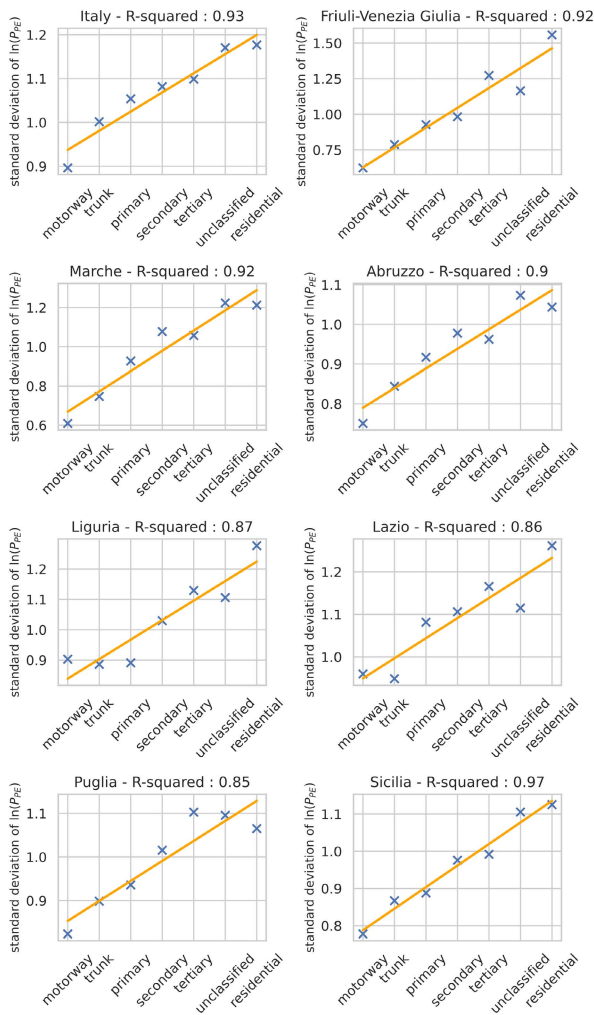


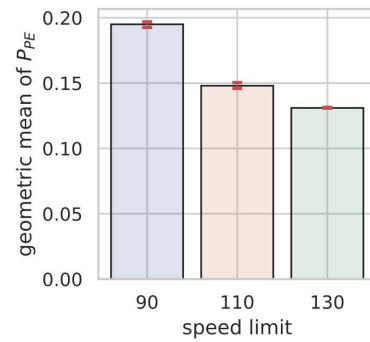
FIGURE 8. Relationship between road ranking and the standard deviation of $\ln(P_{PE})$ in Italy and selected regions.

performed on extra-urban roads where the road categories are characterised by different maximum speed limits and cannot be directly compared. For instance, a speed limit of 90km/h has a different connotation in a primary road and a motorway. The former would have the maximum speed limit allowed for its road type, while the latter would have a significantly lower limit, suggesting differences in road performance, as explained in the following section.

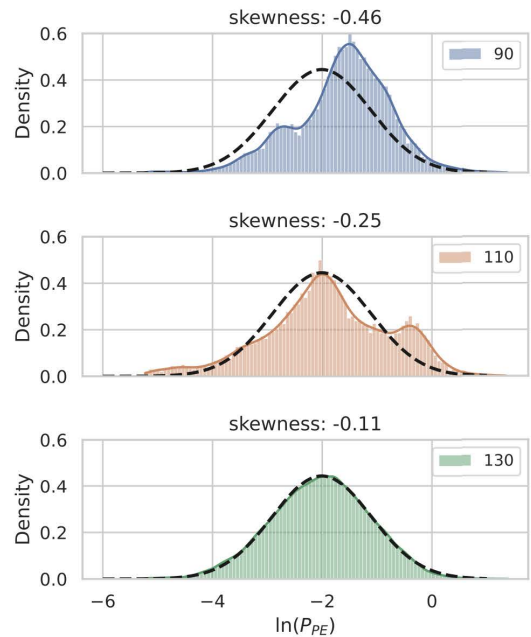
In conclusion, the degree of variability of the road roughness index calculated on a given road is inversely proportional to the importance of that road in the network. This statistic appears to be less sensitive to vehicle speed. It should be regarded as a critical road quality indicator, as it relates to the structural uniformity of the road that strongly reflects on the perceived comfort of the traveller.

5) SPEED LIMIT

The maximum speed travelled on a particular type of road is dictated by each country, and the Italian speed limits currently



(a)



(b)

FIGURE 9. (a) Geometric mean of P_{PE} and (b) distribution of $\ln(P_{PE})$ in motorways at different speed limits.

in force are listed in Table 4. While most streets adopt the recommended speed limit, in many roads, authorities have lowered the speed limit to account for various circumstances (pavement conditions, geographical constraints, number of lanes, level of traffic, presence of bridges or tunnels, curvature of the road). A comparison between the geometric mean of P_{PE} at different speed limits shows that P_{PE} increases as the speed limit decreases in motorways (Fig. 9 a) and other road types. Independent T-test confirmed the differences observed at different speed limits to be significant for each road type (p-value = 0.000). Moreover, $\ln(P_{PE})$ mapped on roads with maximum speed limits distributes normally, whereas roads with a lowered limit show a significant degree of variation from normality (Fig. 9 b), again suggesting the presence of a predominant factor. Most of these routes are in Trentino-Alto Adige and Liguria, mainly due to natural constraints and the presence of many viaducts. In these regions, points with

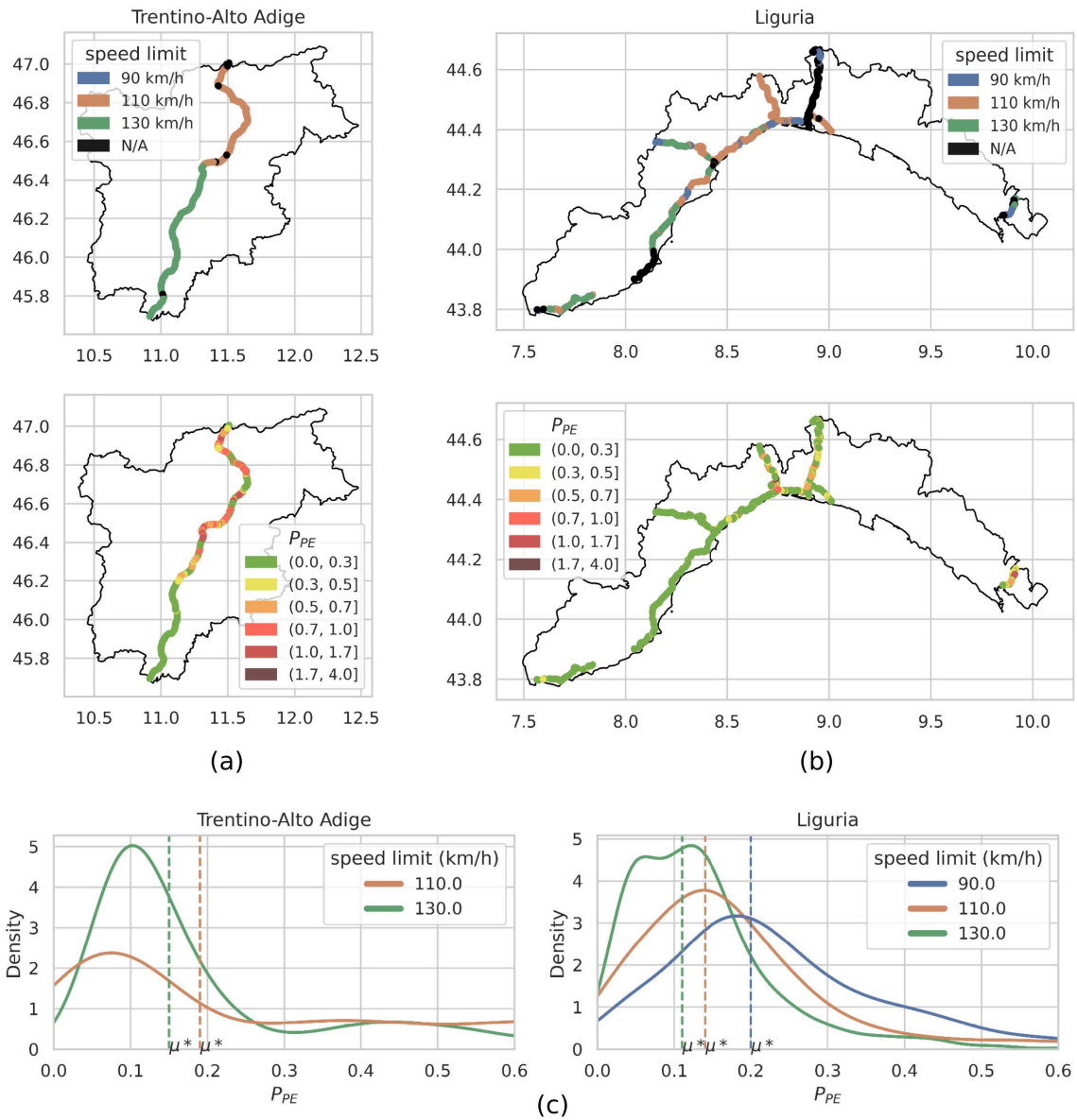


FIGURE 10. Comparison between P_{PE} and speed limit in (a) Trentino-Alto Adige and (b) Liguria. (c) P_{PE} distribution and geometric mean for different speed limits in these regions.

elevated P_{PE} are found almost exclusively on roads with reduced speed limits, while those with maximum speed limit have a consistently low P_{PE} (Fig. 9 a, b, c). A reasonable interpretation for this outcome is in the safety practice of reducing the speed limit when travelling conditions are sub-optimal.

C. COHERENCE WITH PREVIOUS WORK

Previous studies on the relationship between the vehicle speed and P_{PE} showed that the value of P_{PE} depends on the speed of the vehicle according to a gamma law, with parameter $\gamma \in [0, 2]$ and with γ tending to decrease for increasing values of the vehicle speed [30]. The derived coefficient of fitting parameters $P(\nu) = \hat{q} \times \nu^\gamma$ and the corresponding 95%

confidence bounds for motorway roads are $\hat{q} = 16.89 \times 10^{-3} (15.78 \times 10^{-3}, 18.01 \times 10^{-3})$ and $\gamma = 0.76 (0.74, 0.78)$. More precisely, the study suggests that a vehicle travelling on a motorway at a steady speed of 90 km/h and measuring a certain P_{PE} value would produce a value that is 16% higher if it was travelling at 110 km/h, 32% higher if travelling at 130 km/h. The assumption that each vehicle has a stable speed along its journey is reasonable in motorways, where cars travel at the maximum speed allowed. Under this assumption, the recommended gamma law normalisation of P_{PE} was applied in motorways to compare P_{PE} values as if they had been measured at the same vehicle speed. Motorways with a 90 and 110 km/h speed limit, that measured on average a higher P_{PE} than those with a maximum speed limit, recorded

TABLE 4. OSM road categories with corresponding Italian speed limits.

Key	Value	Description	Rank	Type	Speed Limit Urban (km/h)	Speed Limit Extra-Urban (km/h)
highway	motorway	a restricted access major divided highway, normally with 2 or more running lanes plus emergency hard shoulder	1	1	–	130
highway	trunk	the most important roads in a country’s system that aren’t motorways (need not necessarily be a divided highway)	2	1	70	110
highway	primary	the next most important roads in a country’s system (often link larger towns)	3	2	50	90
highway	secondary	the next most important roads in a country’s system (often link towns)	4	2	50	90
highway	tertiary	the next most important roads in a country’s system (often link smaller towns and villages)	5	2	50	90
highway	unclassified	the least important through roads in a country’s system – i.e. minor roads of a lower classification than tertiary (often link villages)	6	2	50	70
highway	residential	roads that serve as access to housing, without the function of connecting settlements (often lined with housing)	7	3	50	–

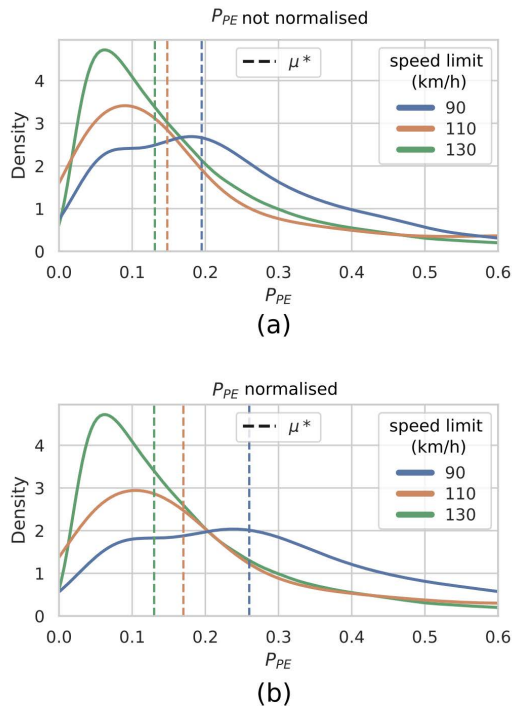


FIGURE 11. Distribution of P_{PE} in motorways (a) before and (b) after vehicle speed normalisation.

an even higher P_{PE} after normalisation (Fig. 11). Thus, the findings on the relationship between the speed limit and P_{PE} persist and are, in fact, amplified when considering previous results of controlled experiments within the SRS project.

V. CONCLUSION AND PROSPECT

The statistical modelling of P_{PE} offers several insights on how to analyse and interpret SRS data. The log-normal nature

of the P_{PE} implies that it should always be described in terms of its geometric mean and standard deviation or its normally distributed transformed variable. The arithmetic mean of the variable misrepresents the sample as it tends to overestimate its central value (due to the AM-GM theorem). All statistical tests assuming additive effects, e.g. linear regression and ANOVA, should be performed in the transformed domain to be valid. The log-normal trend also indicates that a minimal increase in P_{PE} could represent a substantial loss in road performance, suggesting that P_{PE} should be continuously and closely monitored, as already done by SRS. The geometric standard deviation, which defines the shape of the distribution, appears to be the most informative and characterising statistics for this distribution, as observed in other log-normal variables [43]. In fact, the shape parameter was less sensitive to vehicle speed in the study of the effect of road ranking on P_{PE} . Modelling P_{PE} under the law of proportionate effect offers insights into the type of degradation experienced by the road. In the bridge and traffic scenarios, P_{PE} in streets where one or more predominant factors are responsible for the degradation tend to deviate from the expected trend. In contrast, roads with a more homogeneous and “physiological” degradation tend to retain a normal distribution, be it shifted to the right.

The joint statistical analysis of SRS with selected open datasets unveiled the relationship between P_{PE} and various road features. The study was made possible by the open availability of SRS data and greatly facilitated by having P_{PE} values already mapped onto OSM maps. As a result, SRS data can be easily integrated with OSM datasets and the ever-growing number of open-source applications based on OSM. Open data integration is a valuable tool for dataset-wide exploration and should be considered when designing crowdsensing architectures and possibly included

in the data analysis pipeline. In this investigation, the predicted tendency of P_{PE} in relation to bridge, population density, construction materials and road categories was confirmed, providing evidence that P_{PE} reflects road performance. Further proof can be gained by investigating other factors known to negatively affect road performance, such as weather conditions and geographic constraints, upon the availability of suitable network-wide datasets. The study of the combined effect of multiple factors could also support road management by improving road-lifespan estimates and scheduling of road interventions.

REFERENCES

- [1] J. Liu, H. Shen, H. S. Narman, W. Chung, and Z. Lin, "A survey of mobile crowdsensing techniques: A critical component for the Internet of Things," *ACM Trans. Cyber-Phys. Syst.*, vol. 2, no. 3, pp. 1–26, 2018.
- [2] L. C. Klopfenstein, S. Delpriori, P. Polidori, A. Sergiacomi, M. Marcozzi, D. Boardman, P. Parfitt, and A. Bogliolo, "Mobile crowdsensing for road sustainability: Exploitability of publicly-sourced data," *Int. Rev. Appl. Econ.*, vol. 34, no. 5, pp. 650–671, Sep. 2020.
- [3] C. H. Chin, S. Abdullah, S. S. K. Singh, A. K. Ariffin, and D. Schramm, "Durability assessment of suspension coil spring considering the multifractality of road excitations," *Measurement*, vol. 158, Jul. 2020, Art. no. 107697.
- [4] S. Szénási, G. Kertész, I. Felde, and L. Nádai, "Statistical accident analysis supporting the control of autonomous vehicles," *J. Comput. Methods Sci. Eng.*, vol. 21, no. 1, pp. 85–97, Mar. 2021.
- [5] G. Wang, M. Burrow, and G. Ghataora, "Study of the factors affecting road roughness measurement using smartphones," *J. Infrastruct. Syst.*, vol. 26, no. 3, Sep. 2020, Art. no. 04020020.
- [6] M. Y. Shahin, M. I. Darter, and S. D. Kohn, "Development of a pavement maintenance management system. Volume I. Airfield pavement condition rating," *Construct. Eng. Res. Lab. Champaign, IL, USA, Tech. Rep. 45, 1977*. [Online]. Available: <https://apps.dtic.mil/sti/citations/ADA048884>
- [7] J. Y. Wong, *Theory of Ground Vehicles*. Hoboken, NJ, USA: Wiley, 2008.
- [8] T. D. Gillespie, "Fundamentals of vehicle dynamics," SAE Tech. Paper R-114, 1992.
- [9] K. Bogsjö, "Road profile statistics relevant for vehicle fatigue," Ph.D. dissertation, Fac. Eng., Centre for Math. Sci., Lund Univ., Lund, Sweden, 2007.
- [10] T. Lee, C. Chun, and S.-K. Ryu, "Detection of road-surface anomalies using a smartphone camera and accelerometer," *Sensors*, vol. 21, no. 2, p. 561, Jan. 2021.
- [11] K. Chang, J. Chang, and J. Liu, "Detection of pavement distresses using 3D laser scanning technology," in *Computing in Civil Engineering*. Reston, VI, USA: American Society of Civil Engineers, 2005, pp. 1–11.
- [12] R. Gui, X. Xu, D. Zhang, H. Lin, F. Pu, L. He, and M. Cao, "A component decomposition model for 3D laser scanning pavement data based on high-pass filtering and sparse analysis," *Sensors*, vol. 18, no. 7, p. 2294, Jul. 2018.
- [13] M. W. Sayers, "The international road roughness experiment: Establishing correlation and a calibration standard for measurements," *Transp. Res. Inst., Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. 45, 1986*.
- [14] J. Eriksson, L. Giroud, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Proc. 6th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2008, pp. 29–39.
- [15] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. 6th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, 2008, pp. 323–336.
- [16] G. Alessandrini, L. C. Klopfenstein, S. Delpriori, M. Dromedari, G. Luchetti, B. Paolini, A. Seraghiti, E. Lattanzi, V. Freschi, A. Carini, and A. Bogliolo, "SmartRoadSense: Collaborative road surface condition monitoring," in *Proc. UBIKOMM*, 2014, pp. 210–215.
- [17] K. Chen, M. Lu, G. Tan, and J. Wu, "CRSM: Crowdsourcing based road surface monitoring," in *Proc. IEEE 10th Int. Conf. High Perform. Comput. Commun., IEEE Int. Conf. Embedded Ubiquitous Comput.*, Nov. 2013, pp. 2151–2158.
- [18] C. W. Yi, Y. T. Chuang, and C. S. Nian, "Toward crowdsourcing-based road pavement monitoring by mobile sensing technologies," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1905–1917, Aug. 2015.
- [19] A. Mohamed, M. M. M. Fouad, E. Elhari, N. El-Bendary, H. M. Zawbaa, M. Tahoun, and A. E. Hassanien, "RoadMonitor: An intelligent road surface condition monitoring system," in *Intelligent Systems 2014*. Cham, Switzerland: Springer, 2015, pp. 377–387.
- [20] U. Bhatt, S. Mani, E. Xi, and J. Z. Kolter, "Intelligent pothole detection and road condition assessment," 2017, *arXiv:1710.02595*.
- [21] A. Allouch, A. Koubâa, T. Abbes, and A. Ammar, "RoadSense: Smartphone application to estimate road conditions using accelerometer and gyroscope," *IEEE Sensors J.*, vol. 17, no. 13, pp. 4231–4238, Jul. 2017.
- [22] J. Jang, Y. Yang, A. W. Smyth, D. Cavalcanti, and R. Kumar, "Framework of data acquisition and integration for the detection of pavement distress via multiple vehicles," *J. Comput. Civil Eng.*, vol. 31, no. 2, Mar. 2017, Art. no. 04016052.
- [23] D. E. Nunes and V. F. S. Mota, "A participatory sensing framework to classify road surface quality," *J. Internet Services Appl.*, vol. 10, no. 1, pp. 1–16, Dec. 2019.
- [24] M. Ndoye, S. Vanjari, H. Huh, J. Krogmeier, D. Bullock, C. Hedges, and A. Adewunmi, "Sensing and signal processing for a distributed pavement monitoring system," in *Proc. IEEE 12th Digit. Signal Process. Workshop, 4th IEEE Signal Process. Educ. Workshop*, Sep. 2006, pp. 162–167.
- [25] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 247–254, Jun. 1979.
- [26] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*. Boca Raton, FL, USA: CRC Press, 2018.
- [27] V. Freschi, S. Delpriori, L. C. Klopfenstein, E. Lattanzi, G. Luchetti, and A. Bogliolo, "Geospatial data aggregation and reduction in vehicular sensing applications: The case of road surface monitoring," in *Proc. Int. Conf. Connected Vehicles Expo (ICCVE)*, Nov. 2014, pp. 711–716.
- [28] S. Delpriori, V. Freschi, E. Lattanzi, and A. Bogliolo, "Efficient algorithms for accuracy improvement in mobile crowdsensing vehicular applications," in *Proc. UBIKOMM*, 2015, p. 158.
- [29] F. Abbondati, S. A. Biancardo, R. Veropalumbo, and G. Dell'Acqua, "Surface monitoring of road pavements using mobile crowdsensing technology," *Measurement*, vol. 171, Feb. 2021, Art. no. 108763.
- [30] G. Alessandrini, A. Carini, E. Lattanzi, V. Freschi, and A. Bogliolo, "A study on the influence of speed on road roughness sensing: The SmartRoadSense case," *Sensors*, vol. 17, no. 2, p. 305, Feb. 2017.
- [31] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, "A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2419–2465, Apr. 2019.
- [32] L. Cheng, J. Niu, L. Kong, C. Luo, Y. Gu, W. He, and S. K. Das, "Compressive sensing based data quality improvement for crowd-sensing applications," *J. Netw. Comput. Appl.*, vol. 77, pp. 123–134, Jan. 2017.
- [33] B. Wang, L. Kong, L. He, F. Wu, J. Yu, and G. Chen, "I(TS, CS): Detecting faulty location data in mobile crowdsensing," in *Proc. IEEE 38th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2018, pp. 808–817.
- [34] S. Bhattacharjee, N. Ghosh, V. K. Shah, and S. K. Das, "QnQ: Quality and quantity based unified approach for secure and trustworthy mobile crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 200–216, Jan. 2020.
- [35] H. Amintoosi and S. S. Kanhere, "A reputation framework for social participatory sensing systems," *Mobile Netw. Appl.*, vol. 19, no. 1, pp. 88–100, 2014.
- [36] V. Freschi, S. Delpriori, E. Lattanzi, and A. Bogliolo, "Bootstrap based uncertainty propagation for data quality estimation in crowdsensing systems," *IEEE Access*, vol. 5, pp. 1146–1155, 2017.
- [37] M. Castillo, "The wisdom of crowds," *Amer. J. Neuroradiol.*, vol. 34, no. 10, pp. 1863–1865, 2013.
- [38] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Converg., Int. J. Res. New Media Technol.*, vol. 14, no. 1, pp. 75–90, Feb. 2008.
- [39] A. Bogliolo, A. Aldini, G. Alessandrini, A. Carini, S. Delpriori, F. Valerio, L. C. Klopfenstein, E. Lattanzi, G. Luchetti, B. D. Paolini, and A. Seraghiti, "The smartroadsense website," *Tech. Rep.*, 2018. Accessed: Feb. 1, 2021. [Online]. Available: <http://smarroadsense.it>
- [40] *ISTAT open data*, Istituto Nazionale di Statistica, Rome, Italy, 2010. Accessed: Feb. 1, 2021. [Online]. Available: <http://dati.istat.it/>

- [41] *OpenStreetMap Open Data*, OpenStreetMap, Cambridge, U.K., 2007. Accessed: Feb. 1, 2021. [Online]. Available: <https://download.geofabrik.de>
- [42] E. Limpert and W. A. Stahel, "Problems with using the normal distribution—and ways to improve quality and efficiency of data analysis," *PLoS ONE*, vol. 6, no. 7, Jul. 2011, Art. no. e21403.
- [43] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability-normal or log-normal: That is the question," *BioScience*, vol. 51, no. 5, pp. 341–352, 2001.
- [44] J. Eeckhout, "Gibrat's law for (All) cities," *Amer. Econ. Rev.*, vol. 94, no. 5, pp. 1429–1451, Nov. 2004.
- [45] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Math.*, vol. 1, no. 2, pp. 226–251, 2004.
- [46] N. Heckert, J. Filliben, C. Croarkin, B. Hembree, W. Guthrie, P. Tobias, and J. Prinz, "Handbook 151: NIST/SEMATECH e-handbook of statistical methods," NIST Interagency/Internal Rep. (NISTIR), Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 2002. Accessed: May 1, 2021.
- [47] J. N. Yang and S. D. Manning, "A simple second order approximation for stochastic crack growth analysis," *Eng. Fract. Mech.*, vol. 53, no. 5, pp. 677–686, Mar. 1996.



CHRISTEL SIROCCHI received the bachelor's degree in biotechnologies, the master's degree in functional genomics, and the bachelor's degree in applied informatics. She is currently pursuing the Ph.D. degree in research methods in science and technology with the University of Urbino. Her research interests include emergent behavior, collective intelligence, crowdsensing systems, graph theory, and network analysis.



LORENZ CUNO KLOPFENSTEIN is currently a Researcher with the University of Urbino, a Lecturer in applied computer science undergraduate course and the Ph.D. Program in research methods in science and technology, and a Founder and the CTO of DIGIT srl.



ALESSANDRO BOGLIOLO received the Laurea degree in electrical engineering and the Ph.D. degree in electrical engineering and computer science from the University of Bologna, Italy, in 1992 and 1998, respectively. From 1992 to 1999, he was with the Department of Electronics and Computer Science (DEIS), University of Bologna. In 1995 and 1996, he was with the Computer Systems Laboratory (CSL), Stanford University, Stanford, CA, USA. From 1999 to 2002, he was with the Department of Engineering (DI), University of Ferrara, Italy. He joined the University of Urbino, Italy, in 2002. In 2019, he co-founded DIGIT srl, a benefit corporation for digital social innovation. He is currently a Full Professor in computer systems with the University of Urbino. His research interests include mobile crowdsensing, sensor networks, and digital platforms for sustainability and participatory social innovation.

• • •